December 24, 2019

Prediction of credit card defaults

Konstantin Lekomtsev

Springboard Data Science Career Track

Summary:

Objective: choose an optimal classification model allowing to calculate the probability of default and identify main predictors.

Dataset: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Contents

1 OBJECTIVE AND BACKGROUND

According to America Banker as of May 14, 2019 ”Credit card delinquency rates hit a seven-year high in the first quarter largely because many borrowers in their 20s are struggling to keep up with their minimum payments, according to a new report from the Federal Reserve Bank of New York.” [1], Bloomberg as of 9th October 2019 report ”Credit Card Delinquencies in U.S. on Rise for Smaller Issuers” [2] Increased delinquency rates cause a larger number of accounts to become defaulted.

In this project, we aim to choose an optimal classification model allowing us to calculate a probability of a customer default on a credit card and to identify main predictors. We worked with the dataset containing historical data on defaults of the bank customers in Taiwan for the 6 months' period in 2005. Although the details of the financial market in Taiwan back in 2005 and the USA in 2019 might be different, the overall trends are beneficial for understanding more general factors contributing to the event of credit card default. Moreover, as indicated above delinquency rates in the US in 2019 are rising mostly among young borrowers in their 20s and the dataset we have been working with provides a financial health snapshot of a young society where the median age is 35 years old.

2 DATA STORY

We have performed exploratory data analysis of the dataset containing information on the customer age, credit limit, educational level, marital status, sex, as well as bill amounts, payment amounts, and repayment statuses for April through September 2005. The main goal for EDA was to better understand what features are good predictors for a customer's default in October 2005.

We have analyzed the distributions of age, Fig. 1, and credit amount, Fig. 2, for defaulted and not defaulted customers. The median credit amount for the defaulted was 90,000 NT dollars, compared to 150,000 NT dollars for not defaulted. The median age for both categories was 34 years old, however, there was a moderate shift of the distribution's mode for defaulted vs not defaulted customers from 27 to 29 years old. As expected we observed that people who did not default tended to have a larger credit amount and were slightly older.

We then considered the ages, Fig. 3, and the credit amounts, Fig. 4, for the customers who always paid bills on time and those who always delayed payments. These two groups were filtered out based on their repayment status. The customers who always paid on time had a median credit limit of 210,000 NT dollars and those who always delayed had only 70,000 NT dollars. The distributions of credit limit by repayment status are also different, for the customers who always paid on time it is more bell-shaped and less skewed. For the age columns, the median value of age for customers who always paid on time was 36 years old, and for cus-

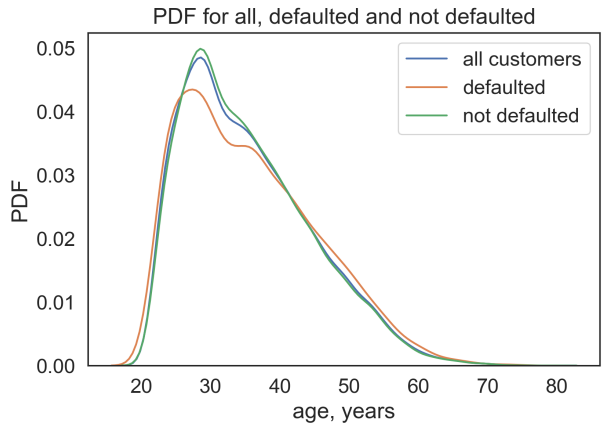

FIGURE 1: CUSTOMER AGES: DEFAULTED VS NOT-DEFAULTED.

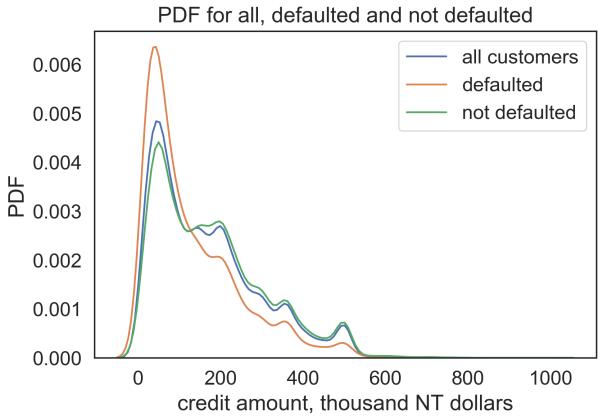

FIGURE 2: CREDIT AMOUNTS: DEFAULTED VS NOT-DEFAULTED.

tomers who always delayed payments it was 34 years old. Similarly to the credit limit distributions, the age distribution for customers who always paid on time is less skewed compared to the distributions for all customers and for the customers who always delayed payments.

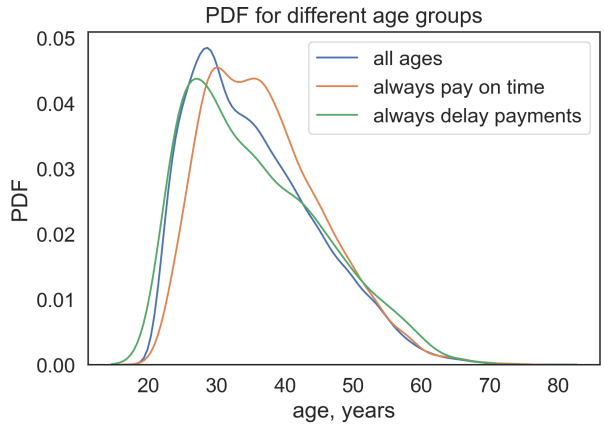

FIGURE 3: CUSTOMER AGES BY REPAYMENT STATUS.

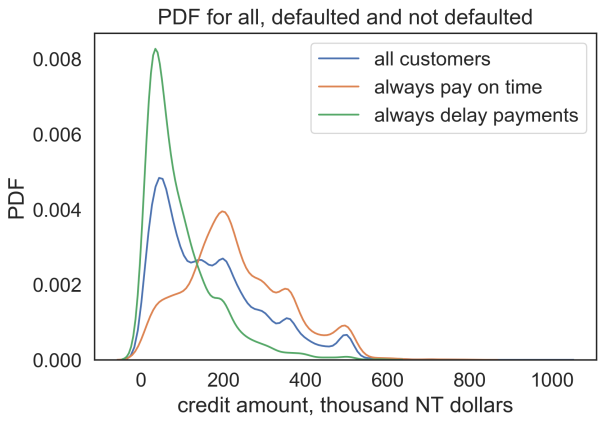

FIGURE 4: CREDIT AMOUNTS BY REPAYMENT STATUS.

From the data on bill amounts and the payment amounts from April through September 2005 we investigated pairwise relationships between bill amount in a given month vs payment amount in the following month using scatter plots, Fig. 5. There are two clear trends: correlation between the minimum payments and the bill amounts, and the full bill repayments and the bill amounts. As expected, the customers who did not default form a larger proportion of those who paid more than a minimum payment each month.

By this dataset, we were given a ”snapshot” of economically active adults in Taiwanese society in 2005. It became clear from basic EDA that we are dealing with a young and educated society with a larger proportion of the female population. 82% of the customers have at least a bachelor’s degree. In each educational category, there are more female customers, which is consistent with the fact that out of all customers 60% percent are female. The majority of cus-

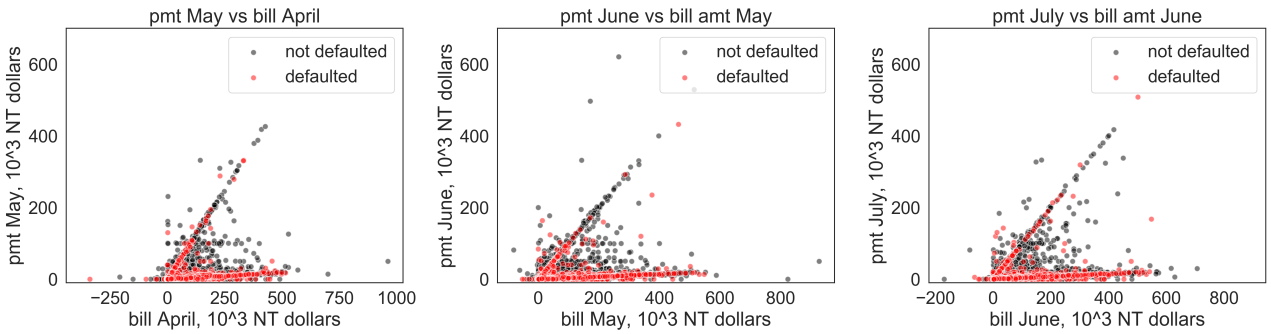

FIGURE 5: BILL VS PAY FOR APRIL TO JUNE.

tomers are either married or single, with a very small proportion divorced. For example, 98.8% of customers with a bachelor's degree are either married or single, the rest are either divorced or classified as other. There are 53.2% single customers in the data set vs 45.53% who are married. Out of those who are married 77.3% have at least a bachelor's degree, out of those who are single 86.6% have at least a bachelor's degree.

3 DATA WRANGLING AND PREPROCESSING.

The dataset is available both on Kaggle and at UCI machine learning repository. It required minimum data wrangling, all features were numerically encoded, so for EDA the columns were remapped from numerical values to strings for easier interpretation and visualisation.

There were outliers in the bill amount and payment amount columns, majority of the outliers showed the customers who paid much larger or smaller amounts compared to the previous month bill, or those who had a negative bill amount. All these situations are expected to occur in practice. The following pre-processing steps were taken before the classification models were applied. The outliers in the numerical columns that were outside

$$lower\,bound = quartile_1 - 1.5 * IQR$$

$$upper\,bound = quartile_3 + 1.5 * IQR$$

were replaced with median values. In this dataset there were duplicate categorical values, we had them remapped. There were no missing numerical values, but as part of pre-processing workflow the functions imputing missing categorical and numerical features were included in the code. To prepare the dataset for machine learning models, the categorical features were converted to dummy variables, the numeric columns were standardised using StandardScaler, which was first used to fit and transform the training data and then transform the test data.

4 ANALYSIS USING MACHINE LEARNING.

4.1 FEATURE ENGINEERING

The dataset contains records of bill amounts, payment amounts, and repayment statuses for 6 months, April, 2005 to September, 2005. We were given the information whether a client defaulted or not in October, 2005. Based on the available financial data additional features were introduced in the dataset. The first set of features was credit utilisation for each month, calculated as

$$credit\,utilisation = \frac{bill}{credit\,amount}.$$

The second set of features was cash flow for each month, calculated as

$$cash\ flow = pmt - bill.$$

The cashflow values are negative when a client pays less than the bill amount for a given month, zero when the bill is fully repaid, and positive when a client pays more than the bill amount.

4.2 CLASSIFICATION MODELS.

We dealt with an imbalanced dataset, where 28% of records are for the clients who defaulted in October 2005. imblearn module was used to produce oversampling only for the training set with the equal number of records for each class by bootstrapping. The model was evaluated on the test set without oversampling. Details available in this jupyter notebook: oversampling approach. As a second approach we used weights adjustment according to class frequencies during training, details here: weighting approach. The dataset was split into training and test sets, validation set was introduced in the training set as a part of stratified K fold cross-validation.

Four different classification models were applied: logistic regression, random forest classifier (RF), RBF kernel support vector machine (SVM) as well as XGBoost classifier. For all models the hyperparameters were tuned, first using validation curves for each parameter and a default classifier, then using grid search cross-validation based on AUC scoring. For the optimized classifiers, generalization performance on the training data was assessed using learning curves for each parameter and stratified 5-fold cross-validation based on AUC score.

The following parameters were tuned, for logistic regression: regularization strength C, for random forest: max depth of trees and number of estimators, for SVM: regularization strength C and kernel parameter gamma, for XGBoost: max depth of a boosted tree estimator, n estimators, learning rate and gamma parameter. Performance on the test data was assessed using the AUC score, ROC, and precision-recall curve.

The best performing classifier was XGBoost using oversampled training data. The corresponding learning curve is shown in, Fig. 6. Cross-validation training AUC score for optimized XGBoost was 0.823 ± 0.008 showing good performance on training data. However, because the test dataset was not oversampled the test AUC score went down to 0.776.

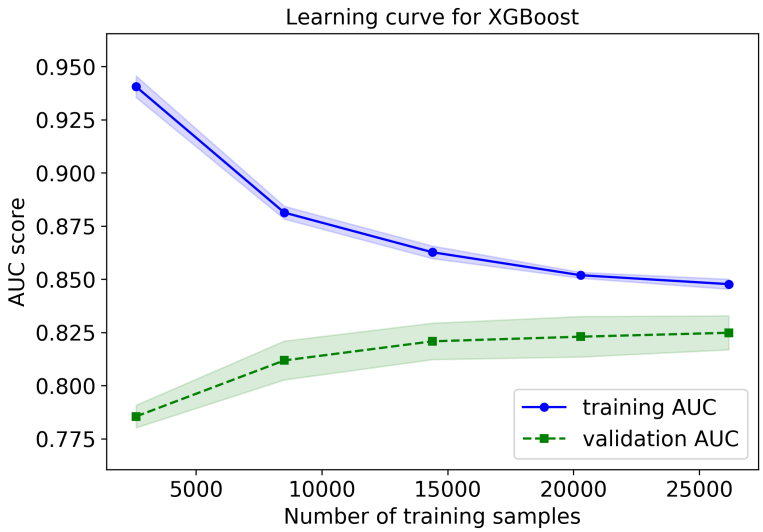

FIGURE 6: XGBOOST LEARNING CURVE.

To further assess the performance of the XGBoost classifier ROC, Fig. 7, and precision-recall curve, Fig. 8, were also computed. By looking at the precision-recall curve we can visually assess that the classifier did not generalize well on the test data. The overall performance of the classifier is shown in Table 1. Poor generalization performance came from a large number of false positives when the negative class (not defaulted) misclassified as the positive class (defaulted), which was a natural consequence of oversampling during training.

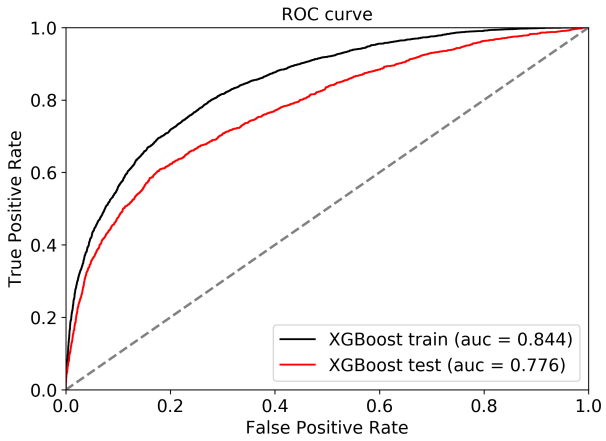

FIGURE 7: XGBOOST ROC.

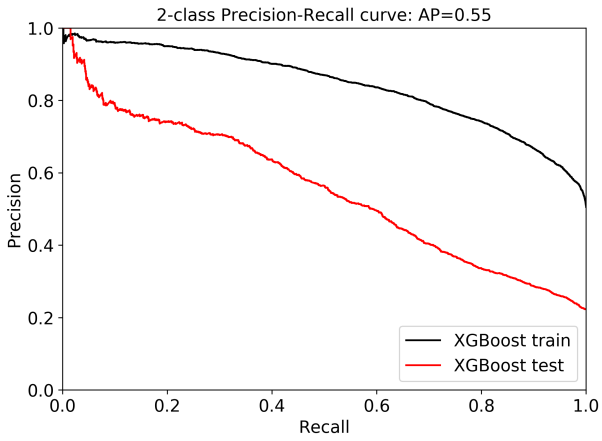

FIGURE 8: XGBOOST PRECISION-RECALL.

XGBoost	precision	recall	f1-score
not defaulted	0.88	0.79	0.82
defaulted	0.47	0.63	0.52

TABLE 1: XGBOOST PERFORMANCE.

A voting classifier combining all four classifiers explored in this study was also applied to the

dataset. However, its performance on the test set was very similar to the XGBoost classifier. Detailed analysis for each classifier is available in the jupyter notebooks accessible by the links that are given at the beginning of this subsection.

4.3 FEATURE IMPORTANCES.

Feature importances were estimated using the optimized RF classifier for the oversampled data as well as for the original dataset. The plots of feature importances are shown in Fig. 9 and Fig. 10.

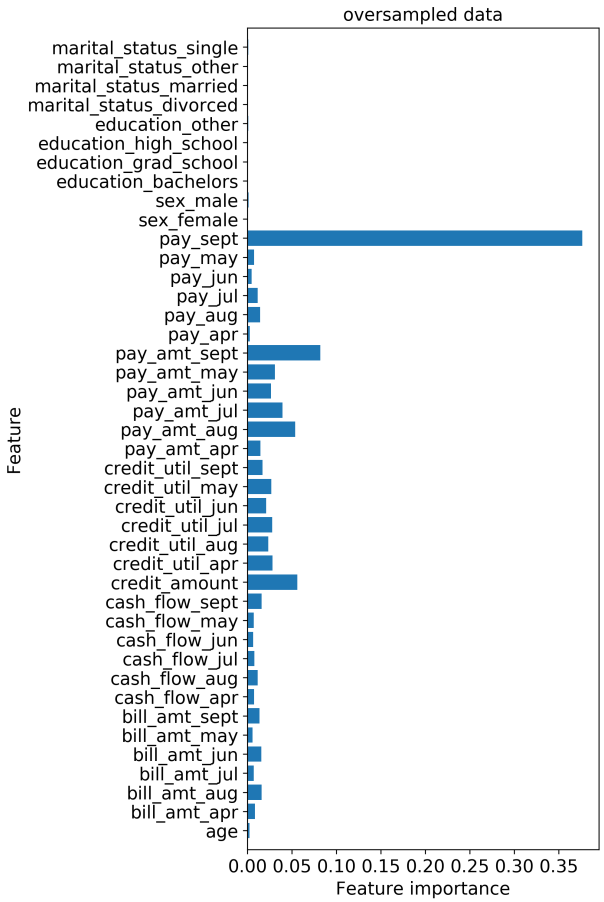

FIGURE 9: FEATURE IMPORTANCE, OVERSAMPLED.

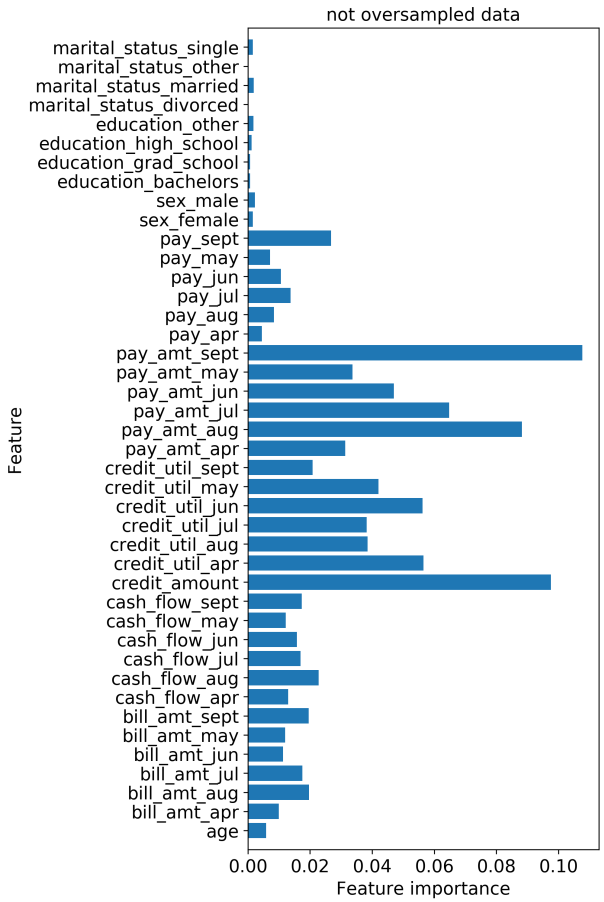

FIGURE 10: FEATURE IMPORTANCE, NO OVERSAMPLING.

In the case of the oversampled dataset, the classifier put a lot more weight on one feature - repayment status in September. In the case of the original dataset, the most important feature is the payment amount in September. The top 5 features in both cases are shown in Table 2 and Table 3. The features containing payment history over the considered 6 months' period are more important compared to the categorical features such as education, marital status, and sex. Repayment statuses and payment amounts 3-4 months before October 2005 are the most important in predicting whether a client defaults or not.

Feature	importance
pay status sept	0.417
pay status aug	0.033
credit amount	0.030
pay amt aug	0.029
pay status jun	0.028

TABLE 2: RF, OVERSAMPLED.

Feature	importance
pay amt sept	0.108
credit amount	0.098
pay amt aug	0.088
pay amt jul	0.065
credit util apr	0.057

TABLE 3: RF, NO OVERSAMPLING.

5 CONCLUSIONS.

We have analyzed a dataset containing payment history and socioeconomic factors for each customer. The goal of the study was to identify the best classification algorithm among logistic regression, SVM, XGBoost and random forest. A voting classifier containing the above four optimized classifiers was also applied to the dataset. An optimized random forest classifier was used to assess feature importances.

The best performing classifier was XGBoost. However, its performance was still sub-optimal for predicting defaults due to a large number of false positives, which was a result of using the oversampled dataset for training. Nevertheless, oversampling yielded better results, compared to putting more weight on the samples of less frequent class during training. One of the ways to improve predictive performance could be adding even more engineered features in the dataset to improve training performance, as getting more data is not possible.

It was identified using the optimized random forest classifier that the repayment statuses and the payment amounts 3-4 months before October 2005 were the most important for predicting defaults.

References

[1] https://www.americanbanker.com/news/whats-behind-the-rise-in-credit-card-delinquencies

[2] https://www.bloomberg.com/news/articles/2019-10-09/credit-card-deliquencies-in-u-s-on-rise-for-smaller-issuers